Sentiment Analysis

A Probabilistic Approach

February 2, 2012

Supervised by I. Langbroek (Blauw Research)
Mentored by dr. M. van Someren (Universiteit van Amsterdam)

S. A. Gieske	S. Laan	C. R. Verschoor	D. S. ten Velthuis	A. J. Wiggers
6167667	6036031	10017321	0577642	6036163

Artificial Intelligence
Faculty of Science
Universiteit van Amsterdam

Contents

1 Introduction

With the growth of Social Media, such as social networks and blogs consumers gained a place to review, rate and recommend products on-line. Making on-line opinion important for companies who want to market their products, manage their reputations and identify new opportunities. Being able to automate the process of finding relevant content, filtering out noise and categorize the messages will speed up the process compared of doing it by hand. Research has shown that sentiment analysis is showing promising results to automate this process.

Sentiment analysis is the application of natural language processing, computational linguistics and text analytics to identify and extract subjective information in source materials. Which makes the data from social media ready for further analysis. Using sentiment analysis we will group twitter messages into positive, negative and neutral messages

2 Problem Specification

Our dataset was provided by our client, a excel file existing of around 10.000 messages from social media[1]. All messages were graded by hand into 5 categories, ranging from very negative to very positive. The goal of our project is to build a learning algorithm that can classify these messages.

2.1 Data Cleaning

A tweet can only contain 140 characters, so to still be able to express yourself users mostly use slang and emoticons. Which increases the value of cleaning the messages first. For example punctuations such as: dot's, comma's and list's dont add anything to the sentiment of a messages so we deleted all useless information. Exclamation and question marks do give value to the sentiment of a sentence because, they can tell something about the emotion of the sentence so we left those in our corpus.

Making use of a Dutch stemmer found in the NLTK database (heb ik nog niet over gehad uitleg nodig), we are able to reduce inflected words to their stem, base or root form. Most words have the same meaning, for example "FUCK and Fucking" are related to each other and by stemming the 2 words to their base form "fuck" we decrease the amount of different words in our corpus and gain more same words (moet beter verwoord worden).

(iets vertellen over tokenizen? dat we een lijst maken van onze woorden?)

2.2 Data Reduction

Because, of the difference between blogs, facebook post and twitter feeds, we decided to only focus on the twitter feeds. Blogs, have larger text than the other two and have

[1]Data contained posts from blogs, posts from facebook and tweets

a good change that different topics and more opinons are discussed than with twitter. Increasing the need for a different aproach for blogs than twitter. Facebook messages look more like twitter but, beacuse of the restriction on the size of twitter messages there is need for a short writting style which is not needed in the other 2 posts.

3 Theory

3.1 Relevant Literature

3.2 Machine Learning

Machine Learning is a branch in the field of artifical intellgence, which looks for a way to develep computer algoritmes cabable of improving predictions or behaviors based on a dataset. A algoritme can take example data to capture the "pattern" in the underlying probability distribution. A succesfull algoritme can automaticly find these complex patterns in "the training data" and make intelligent desisions in new (similar) data.

3.2.1 Artificial Neural Network

An artificial neural network which is a computational model with the same functional aspects of biological neural networks. (plaatje? om uitleg duidelijker te maken) An neural network is build up by a group of connected arifical neurons, which through their connections have an adaptive system that changes its structure based on internal and external information that flows through the network during the learning phase making it possible to find complex patterns in data.

3.2.2 Perceptron

An Perceptron is an example supervised learning algoritme with reinforcemnt. It is a type of artificacial neural network, where their is a one way direction (each neuron in one layer has direct connection to the neurons in the subsequent layer. One learning technique used by an perceptron is the use of back-propagation, where the output values are compared with the correct answer. When there is an error between the output and the correct anwser, the error is fed back through the network. By adjusting the weights of a neuron the algoritme can decrease the margin of the error. This process is repeated until the error is gone or at its smallest. (vertellen hidden layers? miss te ingewikkeld en niet intressant voor ivo)

3.2.3 Maximum Entropy

3.2.4 Naive Bayes Classifier

3.2.5 Weighted Probability Sum

3.3 Classification Measures

In order to test how well a specific algorithm performs, a couple of measurements are used. Each will be explained briefly in terms of table 1.

		Classified Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

Table 1: Table of classification classes

3.3.1 Accuracy

3.3.2 Precision

The precision is defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

3.3.3 Recall

3.3.4 F-Measure

4 Usage and User Guide

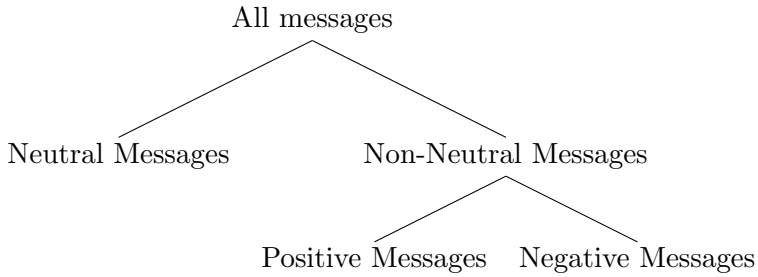
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5 Implementation and Results

5.1 Global Approach

the general approach will be explained first. The idea is to use a binary classifier, i.e. either true or false, multiple times. This way it is possible to first decide whether a message is neutral and then decide between the non-neutral messages whether it is a

positive or negative message. Maybe a tree structure shows the used approach more clearly:



As you can see, using this approach you have to distinguish between two classes at a time. This is very convenient, because there are a lot of techniques that can discriminate between two classes, whereas multi-class-classifiers are less common.

5.2 Binary Classification

In this section, the method that we use to classify a message is described. The approach consists of a few steps:

1. The counting of word frequencies
2. Calculation of word probabilities
3. Calculation of message probabilities
4. Finding a good threshold

First the frequency of each word, i.e. we count how many times it occurs in the training set, is counted. The number The number of times the word occurs in the opposite class, is equivalent to the total number of encounters, minus the encounters in the first class.

When the frequencies of all words[2] are found, the probability for each word can be calculated. This gives an idea of how likely it is to be encountered in a certain class. The formula for this probability is the following:

$$P(word) = \frac{\sum word \in C_1}{\sum word \in C_1 \cup C_2} \tag{1}$$

So the probability that a word is in C_1, the first class, is the number of times it has been encountered in a sentence, which was tagged to be in C_1, divided by the total number of encounters.

Now that all words have probabilities assigned to them, or at least all words in the training set, the probabilities of the sentences can be calculated.

[2] All words in the training set

$$P(s) = \frac{1}{n} \sum_{w \in s} P(w) \tag{2}$$

Where s is the sentence, n the number of words in the sentence and w a word in the sentence. As follows from the formula, the weighted sum of the word probabilities is used.

Now that every sentence has an associated probability, the real machine learning can begin. The goal is to find a threshold for the probabilities; when a sentence probability is higher than a certain amount, it is classified it as C_1, otherwise it belongs to C_2.

5.3 Algoritme 1

5.3.1 Approach

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5.3.2 Results

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5.3.3 Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5.4 Algoritme 2

5.4.1 Approach

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum

ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5.4.2 Results

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5.4.3 Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

6 Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

7 Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.